IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Application of:	)	
	:	Examiner: Not Yet Assigned
BERNARD BREZZO ET AL.	)	
	:	Group Art Unit: Not Yet
Application No.: Not Yet	)	Assigned
Assigned	:	
	)	
Filed: Herewith	:	
	)	
For: SYSTEM AND METHOD FOR	:	
ENABLING A FULL FLOW	)	
CONTROL DOWN TO THE	:	
SUB-PORTS OF A	)	
SWITCH FABRIC	:	June 18, 2001

The Commissioner for Patents
Washington, D.C. 20231

CLAIM TO PRIORITY

Sir:

The applicants hereby claim priority under the International Convention and all rights to which they are entitled under 35 U.S.C. § 119 based upon the following application filed in the European Patent Office:
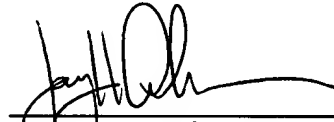
00480054.6 filed June 20, 2000

A certified copy of the priority document is enclosed.

This Page Blank (uspto)

The applicants' undersigned attorney may be reached by telephone at (845) 894-3667. All correspondence should continue to be directed to the address listed below.

Respectfully submitted,

Jay H. Anderson
Attorney for Applicants
Registration No. 38,371

INTERNATIONAL BUSINESS MACHINES CORPORATION
Intellectual Property Law Department
B/300-482
2070 Route 52
Hopewell Junction, New York 12533
Facsimile: (845) 892-6363

this Page Blank (uspto)

Europäisches Patentamt

European Patent Office

Office européen des brevets

Bescheinigung

Certificate

Attestation

Die angehefteten Unterlagen stimmen mit der ursprünglich eingereichten Fassung der auf dem nächsten Blatt bezeichneten europäischen Patentanmeldung überein.

The attached documents are exact copies of the European patent application described on the following page, as originally filed.

Les documents fixés à cette attestation sont conformes à la version initialement déposée de la demande de brevet européen spécifiée à la page suivante.

Patentanmeldung Nr.　　**Patent application No.**　　**Demande de brevet n°**

00480054.6

Der Präsident des Europäischen Patentamts;
Im Auftrag

For the President of the European Patent Office

Le Président de l'Office européen des brevets
p.o.

I.L.C. HATTEN-HECKMAN

DEN HAAG, DEN
THE HAGUE,　　17/10/00
LA HAYE, LE

EPA/EPO/OEB Form　　1014　　- 02.91

This Page Blank (uspto)

Europäisches Patentamt

European Patent Office

Office européen des brevets

Blatt 2 der Bescheinigung
Sheet 2 of the certificate
Page 2 de l'attestation

Anmeldung Nr.:
Application no.:   00480054.6
Demande n°:

Anmeldetag:
Date of filing: 20/06/00
Date de dépôt:

Anmelder:
Applicant(s):
Demandeur(s):
INTERNATIONAL BUSINESS MACHINES CORPORATION

Armonk, NY 10504

UNITED STATES OF AMERICA

Bezeichnung der Erfindung:
Title of the invention:
Titre de l'invention:
System and method for enabling a full flow control down to the sub-ports of a switch fabric

In Anspruch genommene Prioriät(en) / Priority(ies) claimed / Priorité(s) revendiquée(s)

Staat: State: Pays:	Tag: Date: Date:	Aktenzeichen: File no. Numéro de dépôt:

Internationale Patentklassifikation:
International Patent classification:
Classification internationale des brevets:

/

Am Anmeldetag benannte Vertragstaaten:
Contracting states designated at date of filing: AT/BE/CH/CY/DE/DK/ES/FI/FR/GB/GR/IE/IT/LI/LU/MC/NL/PT/SE
Etats contractants désignés lors du depôt:

Bemerkungen:
Remarks:
Remarques:

This Page Blank (uspto)

SYSTEM AND METHOD FOR ENABLING

A FULL FLOW CONTROL

DOWN TO THE SUB-PORTS OF A SWITCH FABRIC

Field of the Invention

5   The present invention relates to communications networks and refers more particularly to the switching nodes of those networks implemented from very high-speed fixed-size packet switch fabrics.

Background of the Invention

10   In recent years, the explosive demand for bandwidth over communications networks has driven developments, some resulting in commercial offerings, of very high-speed switching fabric

FR 9 2000 0028   -1-

devices. The practical implementation of network switching
nodes, capable of handling aggregate data traffic in the range
of hundredths of gigabits per second and soon in terabits per
second, is thus becoming feasible. If many different approaches

5　　are theoretically possible to carry out switching at network
nodes, today's standard solution is to employ, irrespective of
the higher communications protocols actually in use to link the
end-users, fixed-size packet (also referred to as cell) switch-
ing devices. They are simpler and more easily tunable for

10　performances than other solutions especially, those handling
variable-length packets. Thus, NxN switches, which can be
viewed as black boxes with N inputs and N outputs are made
capable of moving fixed-size packets from any incoming link to
any outgoing link. An incoming link is connected to a switch

15　fabric through an input port however indirectly. In practice,
there is a port adapter between the physical incoming link
e.g., a fiber optical connection, and the actual switch fabric
input port in order to adapt the generally complex physical
protocol (and sometimes higher communications protocols too) in

20　use between switching nodes, to the particular switch fabric
input port. Conversely, the interface between the switch fabric
and the outgoing link is referred to as the output port and
there is also an output adapter. Irrespective of how the
switching fabric core is actually devised and implemented this

25　approach is characterized in that the switching fabric itself
does not interface directly to any link external to the switch-
ing node thus, the interface between adapters and switch fabric
along with the corresponding part of the adapter, becomes an
integral part of the switching node and a key parameter to

30　consider for its architecture. Especially, the connections
between the adapters and the switch fabric is an area that
requires careful design. Although, in general, it is preferable
to use parallel connections as much as possible to keep cost
down (since this allows to use slower or current i.e., inexpen-

35　sive chip technologies e.g., CMOS vs. GaAs, for a same through-
put) there is a number of rapidly limiting factors in this

FR 9 2000 0028　　　　　　　　　-2-

direction. Building a very fast switch produces a large number
of I/O connections since there is a multiplying factor i.e.,
the number of ports. A switching fabric is commonly a 16x16 or
32x32 switch thus, with 16 or 32 fully bi-directional ports.

5   Then, parallel connections create a very large number of wires
to be handled both on the backplane and for attaching to the
switch fabric forcing to use expensive module and packaging
solutions. Hence, to push switch performances the other alter-
native is to increase links speed within the limit of the chip

10 technology in use. However, as basic clock speed and number of
wires in each parallel connection both increase, one soon
starts to get problems with skew. That is, the signal on some
paths arrives at a different time from the parallel signal on a
different path. Skew is a very serious limitation to effec-

15 tively being capable of using parallel connections and its
control is a key design issue. Also, to make things worse, the
drivers located at the periphery of chip modules have to be
made slower than those of the interior of the switch fabric
because they have to drive higher value parasitic capacitors

20 requiring switching more current through the parasitic induc-
tances of the packaging and creating a problem known as simul-
taneous switching (ground is disturbed while drivers are
toggling in synch), another drastic limitation to the use of
many signal I/O's.

25   It results from the above considerations that the number
of wires allowed in each port, and the number of ports itself,
of commercially available switch fabrics are a careful tradeoff
between the performances and limitations of the various compo-
nents involved i.e., chip technology, chip packaging (module)

30 technology and board technology along with their respective
costs in an attempt to reach the overall best cost/performance
ratio for a switching node. As a consequence, a state-of-the
art switch is a device having a maximum of a few tenths of
ports e.g., 16 or 32, each having a few data I/O's per port

35 e.g., 4 or 8 for input and the same for outputs (in order it
exists practical solutions to control the skew). Also sometimes

FR 9 2000 0028   -3-

implementing a so-called 2-way data link bundling (two cells
are moved IN and OUT simultaneously). And, since each port is
toggled to the maximum frequency allowed by the current chip
and packaging technologies this allows to match the speed of an

5   OC-192 line i.e., the level 192 of the Synchronous Optical
Network (SONET) US hierarchy i.e., 10 gigabits/s (equivalent to
the European 64th level of the Synchronous Digital Hierarchy or
SDH and called STM-64) over each in and out port yielding to a
128 gigabits/s aggregate throughput switch.

10   On the other hand, another very important item that shapes
the design of switch fabric devices is flow control. A very
simple illustration of the need for a flow control mechanism in
a switch is to observe that when more than one data packet
attempt to access an output port simultaneously (all input

15   ports may want to access the same output port at any given
instant), then a conflict occurs. When this happens, only one
of the contending packet can be read out. Other data packets
have either to be stored in a buffer or queue, until they can
actually be read out, or must be dropped. Although various

20   buffering types are encountered, many of the recent switches
have adopted output-queuing that is, when a packet is arriving
and handled in a switch, it is immediately placed in a queue
that is dedicated to its outgoing port, where it waits until
departing from the switch. This approach is known to maximize

25   the switch throughput provided no input or output is oversub-
scribed. In this case, the switch is able to support the
traffic and the queue occupancies remain bounded. In practice,
output-buffered switches are not however free of complications.
In particular, a NxN switch requires that the internal

30   bandwidth be N times the input bandwidth. Also, the internal
memory space needed in the switch fabric is limited by what
chip technology can reasonably permit (die size, which is by
far the primary contributor setting the cost of a chip, limits
practically the amount of internal memory that can be imple-

35   mented). Then, under unfavorable traffic conditions e.g., with
a high degree of burstiness, the limited on-chip memory

FR 9 2000 0028   -4-

traditionally led to poor throughput especially when FIFO
(First In First Out) input queues used to be utilized at the
input side of the switch fabric i.e., in the input adapter, to
store cells that could not be temporarily accepted by the

5   switch fabric, bound to raise a memory full status. Because
just deploying more on-chip memory to solve the problem is not
economically feasible (even though memory cost has dramatically
dropped over the years) a switch fabric end-to-end traffic
management has thus become an essential piece of a switch

10   design to ensure that no packet are lost, due to congestion and
high utilization while warranting fairness regardless of the
traffic patterns received through the input ports. To this end,
replacing the FIFO queues by VOQs (Virtual Output Queue) in the
input adapter, has contributed to get rid of the well-known HOL

15   (head-of-line) input blocking problem encountered in those of
the switches that are also using input-queuing because VOQ
permits that any packet in a queue, irrespective of its order
of arrival, can be processed provided the individual port
output buffer, to which packet is destined, is not full.

20   However, VOQ mechanism can only work if it has indeed a knowl-
edge of the status of the output buffers i.e., it must know
which ones are full and which ones can still receive cells.
This has necessitated the implementation, in the output
adapter, of an output queue grant-based flow-control mechanism,

25   aimed at passing a grant vector of N bits (one per output over
which the classes of priority handled by the switch can be
time-multiplexed) therefore, at the expense of having to add
more signal I/O's to the switch fabric.

     Much more on switching and switches can be found in the
30   abundant literature that exists on the subject of switch archi-
tecture, their design and limitations and packet switching
networks in general. For example, a good review of switches can
be found in the chapter 5 of a book titled 'ASYNCHRONOUS TRANS-
FER MODE NETWORKS Performance Issues' by Raif O. ONVURAL,

35   Artech House, 1995 and also in a publication by the Interna-
tional Technical Support Organization of IBM, Research Triangle

Park, NC 27709, under the title 'Asynchronous Transfer Mode
(ATM) Technical Overview, no. SG24-4625, October 1995.

Therefore, commercially available fixed-size packet switch
fabrics are carefully crafted to best take advantage of all the
5   capacities of current chip technologies especially, their
intrinsic internal speed, while succeeding in getting around
the limitations imposed by the packaging, characterized by a
scarcity of I/O resources and a drastic limitation in the
number of interconnections that would otherwise ideally be
10 necessary. The result is a piece of hardware having a maximum
of a few tenths of ports (e.g., 16 or 32) however, running at
very high-speed (e.g., OC-192 at 10 Gigabits/second) and
capable of handling the overall full traffic of all ports
without any loss thanks to a sophisticated flow control put in
15 place to manage all sorts of congestion.

It remains however that, in practice, it is often very
difficult to take advantage of the full performance of every
port. Not all the applications require that all ports be of
that speed. On the contrary, many applications of switch
20 fabrics, even though they are attempting to utilize the full
throughput capacity of the switch, rather require that a much
larger number of lower-speed ports be accommodated in a switch-
ing box instead. Switch fabrics are expensive pieces of
hardware. When building boxes, it is then often necessary to
25 combine in the switch fabric port adapters a number of lower
speed lines for obvious economical reasons. For example, a port
adapter, instead of being connected to a single OC-192 line may
have to be connected to four (independent) OC-48 lines each at
2.4 gigabits per second or to sixteen OC-12 lines at 622
30 megabits per second, so as to implement a switching node
comprised of a much larger number of ports, hereafter denomi-
nated sub-ports (since they are derived from a native switch
fabric port) for example implementing, from a 16x16 switch
fabric, a 256x256 switch box concentrating OC-12 lines or any
35 other combination. Unfortunately, switch fabric ports does not

scale down very well because of the sophisticated flow control
mechanisms put in place (in an I/O constrained environment) and
that are thus essentially designed to accommodate a single
high-speed port and are unable to work well if many independent

5 lower-speed line are connected to them instead. To illustrate
this, a port adapter handling e.g., four OC-48 lines has no
mean to report a congestion occurring on a particular path,
while others are not congested. The only solution left with is
to report a global congestion for that port even though 3 lines

10 out of 4 in this case could continue to receive traffic. This
triggers a grossly under-utilization of the capacity of the
port and goes against the objective of trying to take advantage
of the full switch capacity.

Object of the Invention

15 Therefore, it is a broad object of the invention to remedy
the shortcomings of the prior art, as noted here above thus,
keep fully taking advantage of the intrinsic performance of a
N-port switch fabric used to build a M-port switching function
concentrating, through port and sub-port adapters, the traffic

20 of more than N lines.

It is another object of the invention to take into
account the individual traffic of all sub-ports, indirectly
connected to the switch fabric ports, thus enabling an overall
flow control of a switching function irrespective of the physi-

25 cal organization of the core switch fabric in use.

Further objects, features and advantages of the present
invention will become apparent to the ones skilled in the art
upon examination of the following description in reference to
the accompanying drawings. It is intended that any additional

30 advantages be incorporated herein.

Summary of the Invention

A method and a system for enabling a traffic flow control down to all sub-ports of a switching function made of a N-port core switch fabric are disclosed. The switching function

5   comprises one or more port adapters each including one or more sub-port adapters. The invention assumes that, in each sub-port adapter, when a congestion is detected in OUT leg, it is reported through the corresponding IN leg. The detected congestion is piggyback over the incoming traffic entering the

10   input port of the N-port core switching fabric and coming from the IN leg sub-port adapter. Then, in the N-port core switch fabric, the detected congestion is broadcast to all output ports. In turn, in each port adapter, the same information is broadcast to all sub-ports.

15   Then, in each sub-port adapter, a checking of whether OUT leg of a Nth sub-port adapter is reported to be congested or not is performed. If found congested sub-port adapter stops forwarding traffic destined for this Nth sub-port OUT leg and hold further received traffic if any. Sub-port adapter keeps

20   or resumes forwarding traffic, if any received, destined for this Nth sub-port OUT leg as soon as it is reported not to be congested. All sub-port adapter congestion reporting is cycled through and acted on similarly.

Therefore, the invention allows to take advantage of the

25   full intrinsic performance of a N-port switch fabric used to build a M-port switching function concentrating, through port and sub-port adapters, the traffic of more than N independent lines.

Brief Description of the Drawings

Figure 1 discusses prior art and introduces the problem
(a, b, c) solved by the invention.

Figure 2 further discusses the problem solved by the inven-
tion.

Figure 3 is an overall description of the invention.

Figure 4 shows an example of the interface to a switch
fabric port and how sub-port congestions can be
reported.

Figure 5 is a diagram of the steps of the method per the
invention to report congestion from a sub-port.

Figure 6 is a diagram of the steps of the method per the
invention to check if a congestion is reported from
a sub-port.

Figure 7 is a system per the invention.

FR 9 2000 0028   -9-

Detailed Description of the Preferred Embodiment

Figure 1 illustrates prior art showing a high-performance switch fabric of the kind best suited to take advantage of the present invention.

5   **Figure 1-a** is thus a conceptual view of a fixed-size packet switch fabric [100] capable of switching fixed-sized packet (also often referred to as cell) [110]. Packets are comprised of a header [112] and a data part [114] i.e., the payload, each packet transporting a small (fixed-size) chunk

10   of the information exchanged by end-users. Header contains all the necessary information so as packet can properly be handled and steered through the switch fabric [100]. That is, a packet entering switch fabric through an input port [120] exits it through an output port [130]. Ports are paired [140], includ-

15   ing an input and an output port, so as data can flow in both direction along a path linking the end-users, possibly through many equivalent such switch fabrics installed at nodes of a data communications network. Switch fabric [100] shown here, as an example to illustrate the invention, is a 16 x 16 switch

20   fabric. Hence, any packet as [110] entering it through an input port as [120] can be directed to any of the 16 output ports such as [130]. Switch fabric has also the built-in capability of replicating a same input packet e.g., [150]; if instructed to do so in the packet header, over more than one

25   port (up to all 16 output ports) whenever a multi-cast or broadcast is necessary i.e., when a packet needs to be distributed to more than one destination in the network. This is thus exemplified here with a packet entering through input port [150] and replicated over the three output ports [151,

30   152, 153].

Figure 1-b shows that the switch fabric [100] is, in practice, never used alone. Each port pair is connected to a port adapter [160] having an IN [161] and an OUT [162] leg. Port adapter, as the name suggests, is in charge of adapting a

FR 9 2000 0028   -10-

switch port pair on one hand e.g., [170] to a transmission
medium, often a telecommunications line [175] on the other
end. As a typical example of the state of the art, the
telecommunications line is an OC-192 optical fiber line i.e.,

5   corresponding to the level 192 of the Synchronous Optical
Network (SONET) US hierarchy close to 10 gigabits/s (equiva-
lent to the European 64th level of the Synchronous Digital
Hierarchy or SDH and called STM-64). To be able to cope
instantaneously with this steady state speed switch fabric

10   port [170] speed is made even higher and can reach e.g., 16
gigabits/s. Therefore, a switch fabric of the kind shown in
figure 1 is made capable of sustaining an aggregate throughput
of 16 x 2 x 10 or 320 Gigabits per second while being capable
of copping, at switch fabric port, with an IN and OUT instan-

15   taneous throughput of 16 gigabits/s. Therefore, the egress
buffer [164], always present in the output leg [162] of the
adapter [160] may have to fill [166] at an instantaneous rate
of 16 gigabits/s even though it is possibly drained out [168]
at the maximum rate of the line i.e., 10 gigabits/s. It is

20   then subject to overflow especially when many input ports
[172], possibly all, are sending traffic simultaneously to the
same output port for a while.

Hence, switch fabric is typically part of a larger unit
[105] e.g., a switching box to implement a network node, that

25   may comprise up to 16 port-to-line adapters similar to the one
shown [160] in this particular example to illustrate the
invention.

Figure 1-c illustrates with more details one of the chief
problem, briefly suggested here above, encountered with all

30   switching functions and to be solved between switch fabric and
surrounding port adapters. Depending on the traffic character-
istics at a given instant many, if not all, of the input ports
[172] are receiving traffic for a same output port e.g.,
[174]. If the aggregate traffic exceeds, for a significant

35   period of time, what is drained out [168] through the line
[175] connected to the adapter [160] then, egress buffer [164]

FR 9 2000 0028   -11-

eventually overflows and corresponding packets are discarded. Since it exists drastic specifications on the number of packets that can be discarded in a network (all together no more than 1 over 10^9 packets are allowed to be discarded) all

5   modern switches have flow control mechanisms intended to prevent this from ever happening. Namely, whenever OUT leg [162] of a port adapter [160] detects that its internal buffering is near to be exhausted it raises a signal [182] to the switch fabric [100] indicating it can no longer accept

10   incoming traffic for that port. Then, switch traffic must hold what it has already received, in the switch fabric itself, for that port (if switch fabric has indeed provisions to do so). More importantly, it broadcasts [184] to all IN legs of port adapters, such as [191], the information it cannot, in turn,

15   accept traffic for that particular output port [174]. This information is thus used by all the adapters (actually, by all adapters receiving traffic for that particular port) to hold it in their internal buffering generally implemented under the form of a FIFO (first in first out) or with a more sophisti-

20   cated VOQ (virtual output queue) [194], this latter approach allowing to get around the well-known HOL (head of line) blocking observed when FIFO's are used. With a FIFO, when a packet cannot be delivered because the output port it must exit through is busy, all other packets, waiting in line

25   behind, cannot be processed either even though the ports through which they have to exit are idle. VOQ allows to get around this.

The way the overall mechanism to handle congestion from an output port is actually reported and acted on within switch

30   fabric and switching functions varies largely with the numer- ous different implementations found of these functions. This is anyway beyond the scope of the invention which does not rest, as explained later in the description, on a particular mechanism implemented in the switch fabric to be fully effec-

35   tive. Irrespective of the details of a particular solution retained to implement a switch, the idea is anyway always the

FR 9 2000 0028   -12-

same that is, in a switching function [105] all the parties
involved are made aware, through a specific flow control
mechanism, of an output port congestion. Among many alternate
possibilities known from the art, this is simply exemplified
5   in figure 1, through the use of signals [182, 184] raised
respectively to switch fabric and, through switch fabric,
broadcast to all port adapters to inform them of a congestion
occurring at some of the port outputs. The objective is to be
able to use the switching function internal buffering [105],
10   including switch fabric [100] (if switch fabric has indeed
provision for temporarily holding packets i.e., if it is more
than just a switch matrix or crossbar) and port adapter IN and
OUT buffering [164, 194], to their full extent, in an attempt
to prevent any discarding from happening or to delay this
15   event as much as possible this, without impairing traffic of
non congested ports. In an even more global approach of
solving congestion in a communications network it is worth
noting here that some communications protocols may handle
further this mechanism by permitting that the remote source of
20   data be eventually inform of slowing down in case of severe
and/or long congestion. This is for the example the case of
ATM (Asynchronous Transfer Mode) networks implementing an
adaptive flow control mechanism known under the name of ABR
(Available Bit Rate), a service specified by the ATM Forum
25   Traffic Management Sub-working Group. In which case one of a
role of an adapter such as [160] is to inform the remote
source, through the appropriate mechanism of the protocol in
use, it has to pace the sending of data to prevent discarding.

Figure 2 discusses the problem, solved by the invention, and which arises when an adapter [260] is designed to interface more than a single communication line. Although any number of slower lines may have to be handled, 4 or 16 lines
5   (whose aggregate throughput must stay within the one of a single line discussed in previous figure though) are typical examples of what may be needed in actual implementation. Four sub-port adapters [210, 220, 230, 240] are then used in this example to illustrate the problem. To be more specific if
10   communication line [175] of figure 1 was an OC-192 optical fiber line at 10 Gigabits/second then the four lines [215, 225, 235, 245] are e.g., OC-48 lines at 2.48 Gigabits/sec each. In practice, very often, due to the fact that switch fabric are very high performance pieces of hardware imple-
15   mented in an I/O constrained packaging environment, as discussed in the background section, a switch fabric port as [270] has far too much performance (16 Gigabits/second was assumed in Figure 1) to accommodate a single communication line e.g., [215]. Therefore, in order to take full benefit of
20   this performance, lower speed lines are grouped on a same port adapter [260] so as to keep the switching unit cost perform-ance ratio competitive. However, this creates a very serious problem since there is now more than a single line on the same port adapter and only one path [282] for reporting congestion
25   to and through the switch fabric as explained in Figure 1. The straightforward solution to overcome this, results in poor performance. If the filling of the four egress buffers [214, 224, 234, 244] is OR'ed [283] to report a congestion the obvious consequence is that any congestion affecting a line
30   prevents all the other lines from being able of forwarding any traffic at all. Therefore, the objective of allowing to inter-face four independent lower-speed lines, in this particular example, through a single full-speed port [270] is not met since the lines are not really independent. Even if a more
35   sophisticated approach is sometimes considered in which a single egress buffer is maintained for the four lines, so as

to share dynamically a global larger resource between the four
lines thus, attributing a larger share to a line when neces-
sary; this can only delay the occurrence of the problem in
case of a long congestion on one line. Moreover, because the
5   speed of the switch fabric port [270] stays the same i.e., 16
Gigabits/sec in this example, it remains that, for a while,
all traffic can have a same sub-port and lower-speed line as
target thus, exacerbating the problem. Therefore, such a
switch fabric does not really scale down while it should
10   ideally permit, as a building block, to construct any box not
only concentrating traffic solely from the higher-speed lines
it can accommodate but as well from many more lower-speed
lines, when required to fulfill the specifications of a
particular application, yet permitting flow control independ-
15   ently over each of those lower-speed lines and sub-ports.

     Figure 3 depicts the principle of the solution brought by
the invention to this problem. When too much traffic converges
towards an OUT leg e.g., [322], of a sub-port adapter [320]
packets received through the IN leg [321], entering switch
20   fabric through input port [372], are piggy back with the
information that corresponding OUT leg is becoming congested.
The information, contained in the header of each entering
packet [310], is then broadcast [330] within the switch fabric
[300] core of the switching function [305] to all its output
25   ports through the same means, that is, all packets exiting
switch fabric output ports start carrying the information that
OUT leg of sub-port adapter [320] in global adapter [360] of
switch fabric port [370] is becoming congested. In turn all
sub-port adapters, of all global adapters such as [390] are
30   thus updated [385] with the same information. Consequently,
all entities that may have to forward traffic to the congested
sub-port OUT leg [322], are thus made aware of the fact that
sub-port is indeed congested so they should withhold the
sending of more data to this direction. It is worth noting
35   that each sub-port adapter e.g., [340], part of the global

FR 9 2000 0028   -15-

adapter [360] from which the congestion is reported (by
sub-port adapter [320]) are made aware through the exact same
mechanism even though they are located on the same switch
fabric adapter and could be informed directly however, at the
5   expense of uselessly introducing a different mechanism for
reporting congestion.

Although, the mechanism of the invention is mainly
discussed around global adapters such as [390] and [360], each
implementing four sub-ports, it must be clear to those skilled
10   in the art that any number of such sub-ports can potentially
be accommodated, as shown with [398] and [399], while their
aggregate throughput should stay below the one supported by a
switch fabric port.

Also, it must be understood that the invention still
15   apply even though not all adapters have multiple sub-ports. In
other words, the invention works as well in the general case
where some of the port adapter are single sub-port as shown
with [397].

Figure 4 describes, through an example of a preferred
20   embodiment of the invention, the transport of the flow control
information within a switch fabric. It is assumed in this
example that switch port are operated in a so-called two-way
link bundling mode so that, over each port, two 64-byte
packets [400, 410] are processed simultaneously in order to
25   obtain the required level of performance assumed in this
description of the invention i.e., 16 gigabits/sec for each IN
and OUT port. Then, each IN or OUT port is actually made of 8
individual links [420], indexed from 0 to 7, each capable of
toggling at a rate of 2 gigabits/second, the higher rate that
30   can be accommodated with the current packaging and chip
technologies in use. Therefore, two 64-byte packets are trans-
ferred over 8 links in 16 one-byte transfers [430]. Each
packet has its header part [440, 450]. One byte [451] being
devoted to the transfer of the flow control information down

to the sub-ports. Since a byte is insufficient to transport
the flow control information about all sub-ports this latter
is time multiplexed over a continuous set of packets. It is
worth noting here that ports are never actually idle. Even
5   though there are no data to be transferred over a particular
port idle packets [402] are transferred instead of data
packets [404]. Among other things, idle packets are useful to
keep these very high-speed links in synch and the header bytes
like [451] can keep moving the information necessary to
10 properly operate the switch and adapters such as the flow
control herein discussed. Thus, depending on the number of
ports of the switch fabric and the number of sub-ports to be
supported in a specific application, among many possible
alternate solutions, a flywheel mechanism [460] is put in
15 place so as, over a contiguous set of packets, each individual
participant (i.e., the switch fabric as a whole and all the
adapters down to the sub-ports) is kept updated of the conges-
tion status of all the other actors. Therefore, flywheel
cycles through every port and sub-port [461] and possibly
20 through every traffic priority class [462] supported by the
switching function (most of the time classes of traffic are
also supported in order to give precedence to priority flows
and start discarding lower priorities first in case of conges-
tion). The only assumption on which invention rests is that
25 the switch fabric is indeed capable of performing internally a
broadcast of the flow control especially, byte [451] in this
particular example, from any switch fabric port to any other
switch fabric port (as shown in figure 3) so as all adapters
and sub-port adapter can actually be updated.

30   **Figure 5** shows the steps of the method per the invention.
in a sub-port adapter. When there is slot in the flywheel for
the sub-port considered then, OUT leg is checked [510] to
determine if it is congested according to whatever criterion
has been retained for that purpose. If answer is positive
35 [511] the corresponding congestion bit is set in the current

FR 9 2000 0028   -17-

IN packet [520] ready to enter the switch fabric from the IN
leg. This may be a true data packet (i.e., carrying end-user
data) or just an IDLE packet if there nothing to send. If no
congestion is detected congestion bit is reset. This informa-
5 tion is broadcast [530] first to all switch fabric ports
(within the switch fabric) and from all output ports to all
sub-ports [540] in every global adapter. Thus, all sub-ports
are eventually made aware of a congestion that has occurred in
the OUT leg of a particular sub-port adapter.

10   **Figure** 6 further describes the method per the invention
showing that congestion status concerning every OUT leg
sub-port adapter is reported and checked [610] in turn,
depending upon what slot flywheel [600] delivers. If the
reported sub-port OUT leg is indeed congested [611] then,
15 sub-port adapter in which checking is performed must stop
[621] forwarding traffic to the congested sub-port thus, may
have to hold [623) the traffic it has for that destination and
hold further traffic if any is received. On the contrary, if
destination is not congested [612] one keeps forwarding
20 traffic [622] if any is received. All sub-ports are kept
cycling through [630].

   Figure 7 shows a system per the invention using a NxN
port core switch fabric [700] i.e., having N IN and OUT ports
as [710], and allowing to expand it into a MxM switching unit
25 [730] (with M larger N) yet permitting to enforce flow control
down to the sub-ports, such as [720], so that a greater number
of slower ports can be implemented, without having to compro-
mise, from a very high-speed switch fabric used as a building
block.

30

Claims:

What is claimed is:

1. A method for enabling a traffic flow control down to all
sub-ports of a switching function [305] made of a N-port core
5 switch fabric [300], said switching function comprising one or
more port adapters [360], each said port adapter including one
or more sub-port adapters [320], said method comprising the
steps of:

in each said sub-port adapter:

10   detecting a congestion in an OUT leg [322] of said sub-port
        adapter;

        reporting said detected congestion through an IN leg [321]
        of said sub-port adapter [320]; said step of reporting
        further including the step of:

15   piggyback conveying said detected congestion over an
            incoming traffic [310] entering an input port [372] of
            said N-port core switching fabric from said IN leg of
            said sub-port adapter;

        in said N-port core switch fabric [300]:

20   broadcasting [330] said detected congestion to all output
        ports;

        in each said port adapter [390];

        broadcasting [385] said detected congestion to all
        sub-ports;

25 thereby informing all said sub-port adapters of a said
    detected congestion in any one of a said OUT leg.

2. The method according to claim 1 further comprising the steps of:

in each said sub-port adapter:

    checking [610] whether said OUT leg of a Nth sub-port
5   adapter is reported to be congested or not;

    if congested [611]:

        stop forwarding [621] traffic destined for said OUT leg
        of said Nth sub-port adapter, said stopping step further
        comprising the step of:

10   holding traffic [623], in said sub-port adapter, if
            any is received;

    if not congested [612]:

        keep or resume forwarding traffic [622], if any received,
        destined for said OUT leg of said sub-port adapter;

15 keep cycling [630] through each reported [600] said sub-port
    adapter repeating all here above steps.

3. The method according to any one of the previous claims
wherein said N-port core switch fabric is switching fixed-size
packets [400].

20 **4.** The method according to any one of the previous claims
wherein said fixed-size packets, moved through the ports of
said N-port core switch fabric, include fixed-size idle
packets [410].

5. The method according to any one of the previous claims
25 wherein more than a single said fixed-size packet are moved
simultaneously [400, 410] through each port of said N-port
core switch fabric.

6. The method according to any one of the previous claims
wherein the step of piggyback conveying said detected conges-
tion is performed in an header field [451] of said fixed-size
packets.

5 **7.** The method according to any one of the previous claims
wherein the step of piggyback conveying said detected conges-
tion over said incoming traffic is carried out including said
fixed-size idle packet [410].

8. The method according to any one of the previous claims
10 wherein the step of reporting said detected congestion of all
said sub-port adapters is time multiplexed [460] in said
header field [451].

9. The method according to any one of the previous claims
wherein the reporting step includes reporting per priority
15 class [462].

10.A system [730], in particular a switching system expanding
the number of ports of a switch fabric [700], comprising means
adapted for carrying out the method according to any one of
the previous claims.

20 **11.**A computer-like readable medium comprising instructions for
carrying out the method according to any one of the claims 1
to 9.

This Page Blank (uspto)

SYSTEM AND METHOD FOR ENABLING
A FULL FLOW CONTROL
DOWN TO THE SUB-PORTS OF A SWITCH FABRIC

Abstract

5   The invention permits that a traffic flow control, down
to all sub-ports of a switch made of a N-port core switch
fabric, be effective. The switch has ports and sub-ports.
Sub-ports concentrate traffic from lower-speed lines to a
switch fabric native port. The invention assumes that, in each
10   sub-port adapter, when a congestion is detected in OUT leg, it
is reported through the corresponding IN leg. The detected
congestion is piggyback over the incoming traffic entering the
input port of the N-port core switching fabric and broadcast
so that all sub-ports become eventually aware of a detected
15   congestion in any of the sub-ports. Then, each sub-port
adapter performs a checking of the congestion status of all
the other sub-ports and acts accordingly that is, stops
forwarding received traffic destined for congested sub-ports
and holds further received traffic if any until sub-ports are
20   reported to be no longer congested. Or just keeps forwarding
traffic if no congestion is reported.

        Therefore, the invention allows to take advantage of the
full intrinsic performance of a N-port switch fabric used as a
building block for an M-port switching function concentrating,
25   through port and sub-port adapters, the traffic of more than N
independent lines.

        Figure 3.

Prior Art

[152]

[153]

[150]

[151]

[114]

[112]

H　　DATA

[130]

[110]

[120]

[140]

Figure 1-a

[100]

[105]

[172]

[170]

[160]

[161]

[175]

[166]　　　　[168]
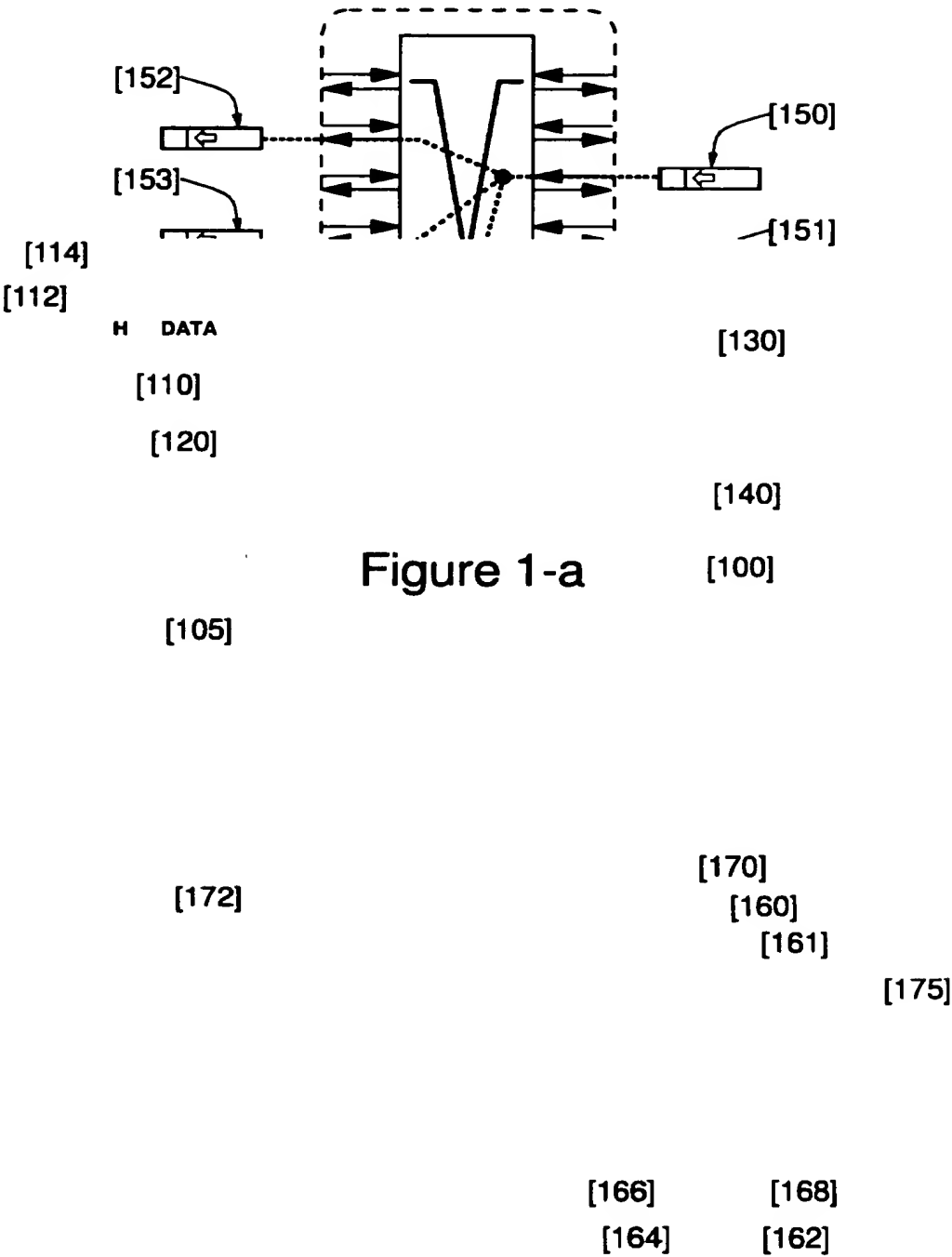
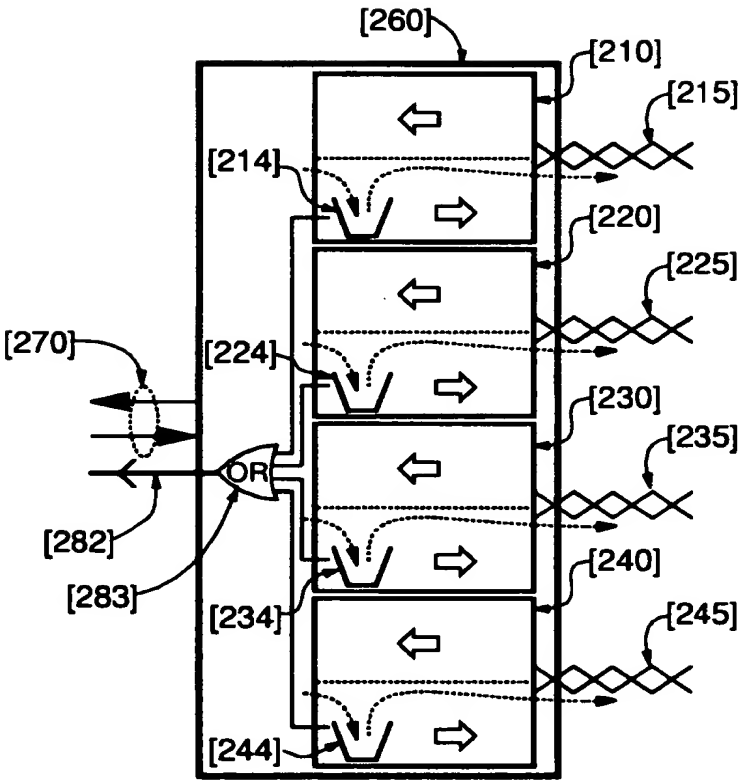[164]　　　　[162]

Figure 1-b

Prior Art
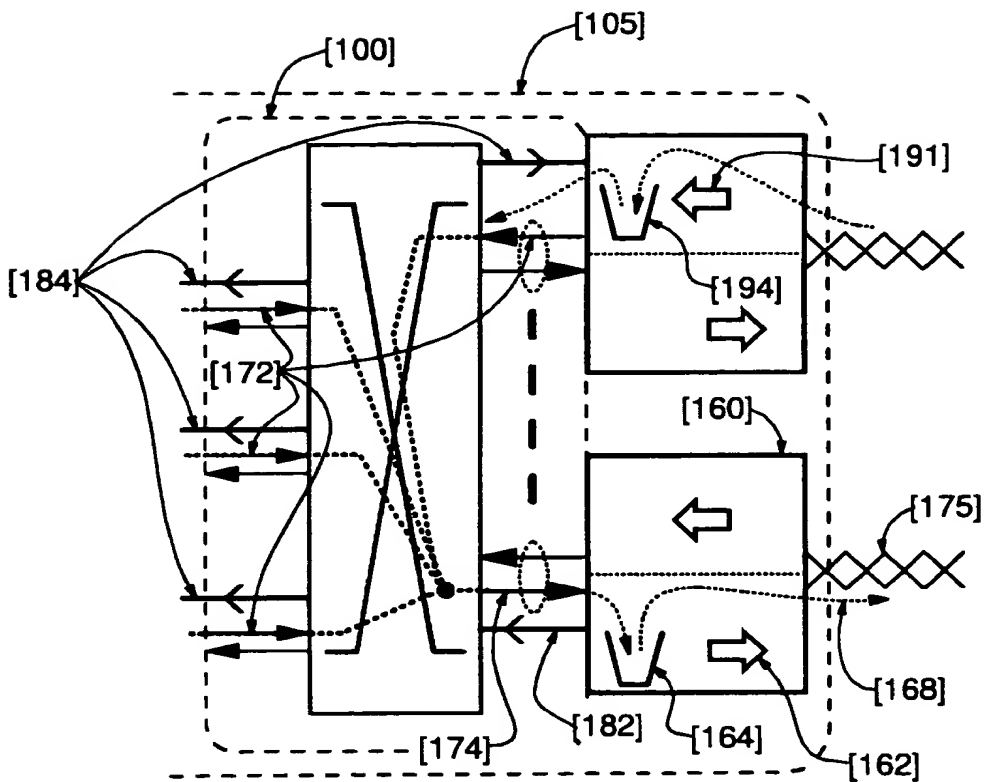
Figure 1-c

Figure 2
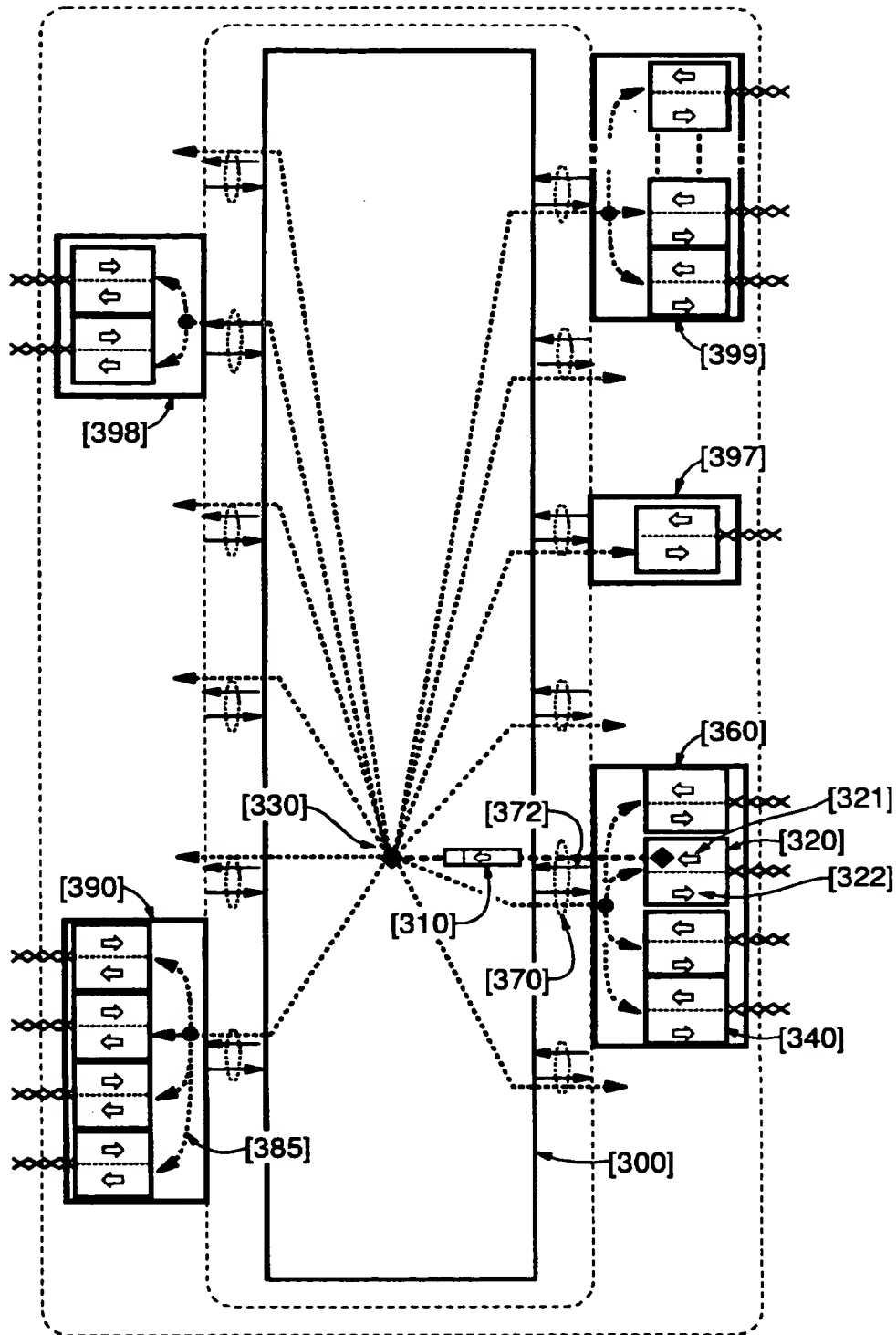
Figure 3

FR 9 2000 0028
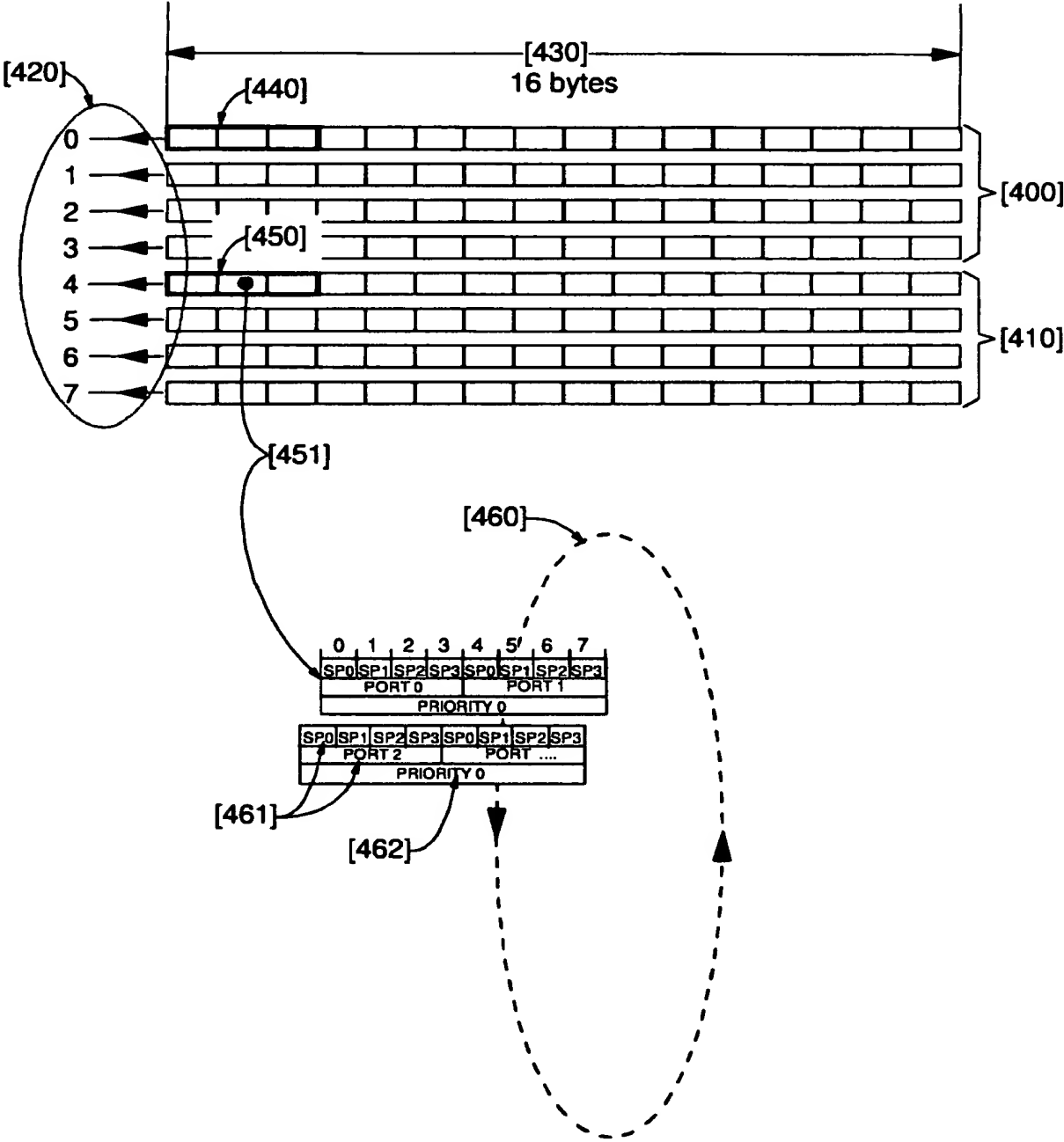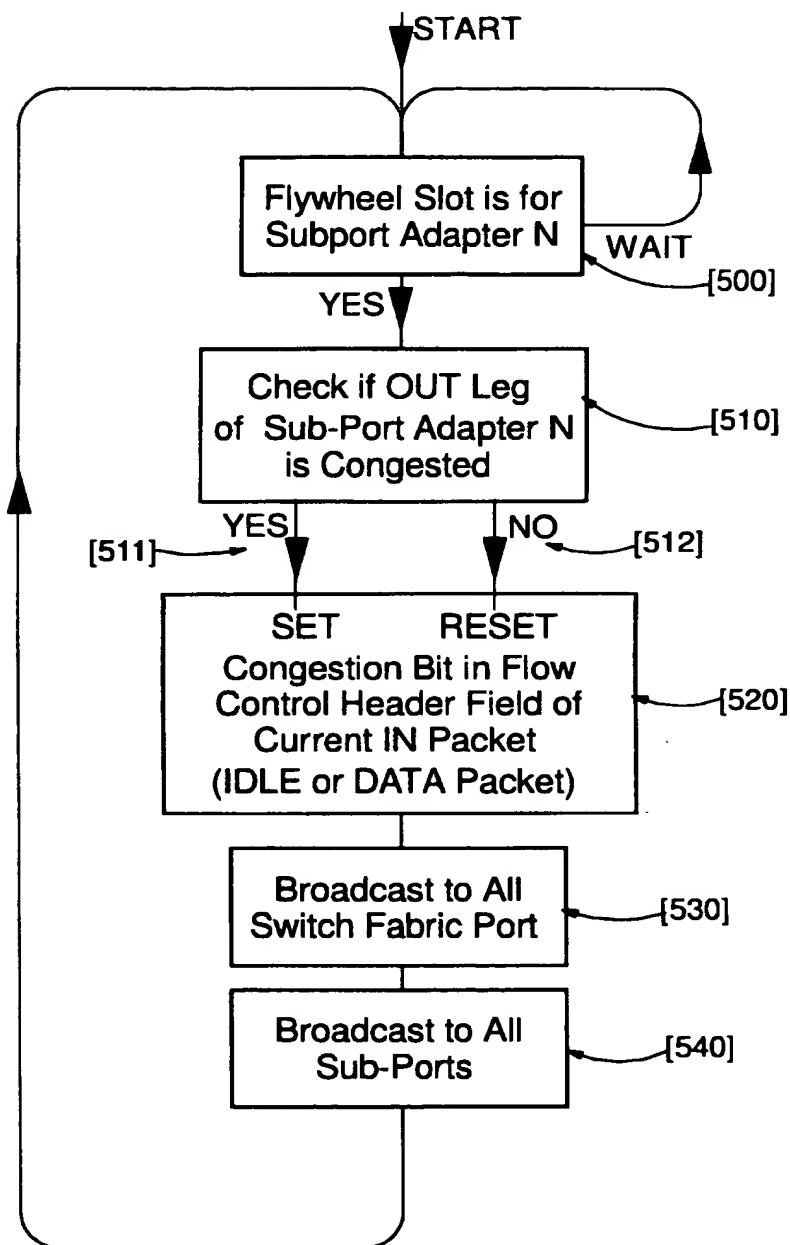**BREZZO et al.**
5/7

Figure 4

FR 9 2000 0028
BREZZO et al.
6/7

START

Flywheel Slot is for
Subport Adapter N   WAIT   [500]

YES

Check if OUT Leg
of Sub-Port Adapter N   [510]
is Congested

YES [511]   NO [512]

SET   RESET
Congestion Bit in Flow
Control Header Field of   [520]
Current IN Packet
(IDLE or DATA Packet)

Broadcast to All
Switch Fabric Port   [530]

Broadcast to All
Sub-Ports   [540]

Figure 5

FR 9 2000 0028
BREZZO et al.
7/7

START

Flywheel Slot is for
Subport Adapter N   [600]

Check if OUT Leg of
Sub-Port Adapter N   [610]
is Reported to be Congested

[611] YES   NO [612]

STOP   KEEP
[621] Forwarding Traffic   Forwarding Traffic [622]
to Sub-Port N   to Sub-Port N
(IF Any)

[623] Hold Traffic
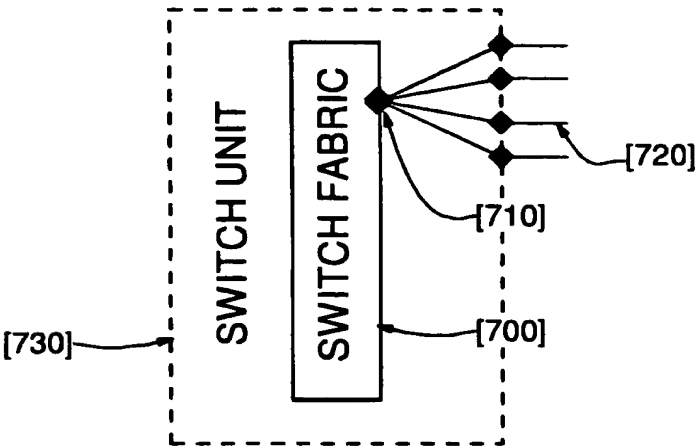(If Any)

[630]

Figure 6

SWITCH UNIT

SWITCH FABRIC

[720]

[710]

[730]   [700]
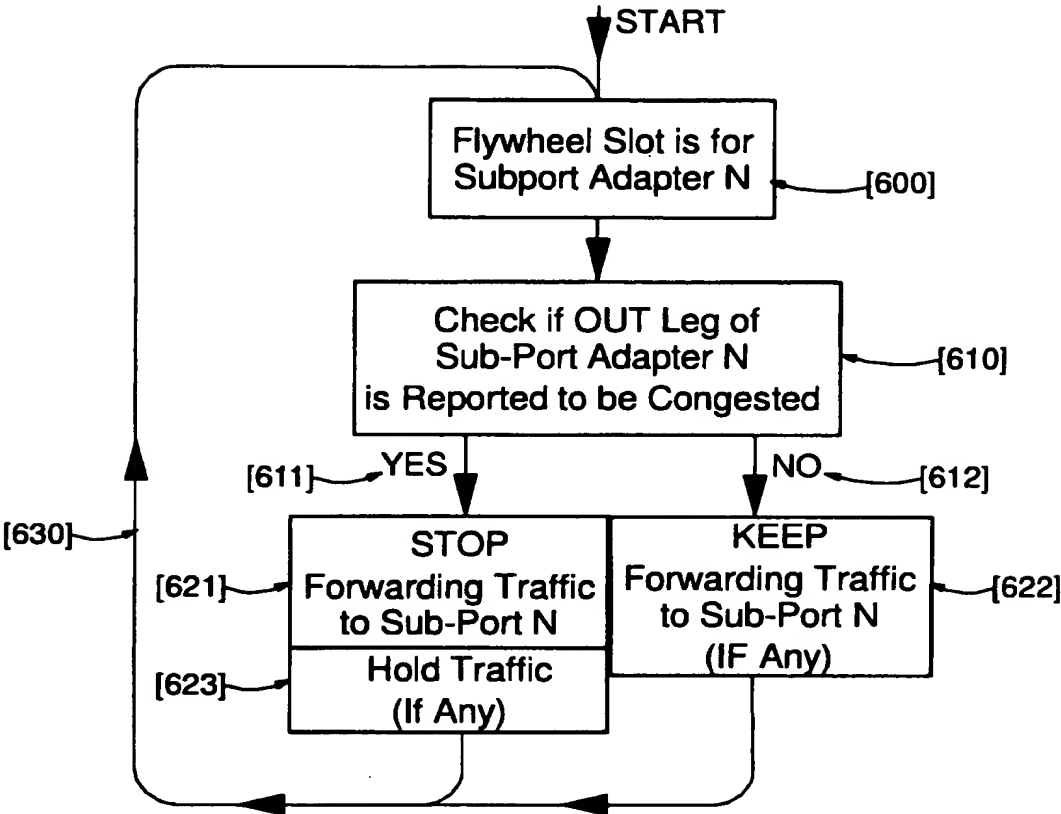
Figure 7

This Page Blank (uspto)